Research Methodology for Educational Data Mining in India

Nidhi Chopra\textsuperscript{1} and Manohar Lal\textsuperscript{2}
\textsuperscript{1,2}School of Computer and Information Sciences, Indira Gandhi National Open University, Maidan Garhi, New Delhi – 110068, India, \textsuperscript{1}nidhi.chopra@gmail.com and \textsuperscript{2}mlal@ignou.ac.in

ABSTRACT
As the world around is going through a technological revolution with the dawn of digital age, we are in some ways compelled to rethink our education system and its components. With the tools and the techniques available to us nowadays it’s imperative for us to reconsider how we can use those to improve our education system. Opportunities for knowledge discovery in educational data have increased tremendously now as compared to the scenario a few years ago. Educational data is becoming increasingly rich as more and more educational systems are going online and collecting large amounts of data. In this paper we will present a study on developing a research methodology for educational data mining. We will focus specifically on the Indian education system and compare it to another university data in USA. Opportunities that are available to us for making effective conclusions using educational data mining are discussed as well.

KEYWORDS
Educational Data Mining, Research Methodology, Indian Education System.

1. INTRODUCTION
In this new digital age, the world of education has also undergone a major transformation. The new technologies and gadgets available help us not only enrich and enhance our existing education system but also offer new opportunities and modes which can take the process of learning beyond institutions and allow people to learn on their own time and own terms. These new advances in learning have played a big role in this age of knowledge enhancement via different means and are clearly a sign that we need to rethink, how we can tap into the technology potential to improve our education system \cite{1}. As of now most of the changes can be seen in the way information is distributed or provided to the students such as e-learning, distance learning, blended learning etc. Another area where we are beginning to see an impact is educational data mining (EDM).

The knowledge revolution has already transformed most work practices and professional jobs. Many jobs which were work intensive before have become more knowledge intensive now. The nature of work has changed from being task oriented to inferential and abstract oriented. Even if you consider a secretarial job for an example, where typically in past that would have involved typing documents or memos, now that involves handling interactions with people inside and outside the corporation. People in all fields and disciplines are becoming more and more informed and are learning to observe, collect and interpret data trends around them to make better and informed decisions. Hence, there is no reason field of education should be any different or left behind \cite{1}.

2. EDUCATIONAL DATA MINING
In the last few years EDM has emerged as a field of its own. The EDM community website, www.educationaldatamining.org, defines EDM as follows: “EDM is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in. \cite{2}” Data mining (DM), or Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data \cite{3}. Largely it consists of analyzing available sets of data to interpret and isolate the trends and patterns present in the data i.e. converting raw data into information so that it can be used by educators, educational software developers, teachers, parents or students. However, it is largely understood that EDM methods are often different from standard DM methods. This is because of the non-independence and multilevel hierarchy found in educational data \cite{4}. Actually because of this reason, it is increasingly common to see the psychometrics models being used in EDM.

![DM as a confluence of multiple disciplines](image)

Figure 1 above shows how DM can be visualized as a confluence of multiple disciplines. In figure 1 the area of study would be education. The data can be collected from students’ use of interactive learning environments, computer-supported collaborative learning, or administrative data from schools and universities.
There are various challenges in the upcoming field of education like understanding choice of major, student drop out, retention, assessment etc. These problems can be solved by using prediction methods [3] of DM. DM is a field which has originated from databases and Artificial Intelligence (AI). Understanding the current trends of our education system and society would clearly point out towards the underlying issues and help us devise an effective plan to address them. Figure 2 shows the two possible dimensions of EDM.

Fig2: Two possible dimensions of Educational Data Mining research.

3. EDM FOR EDUCATORS

An excellent example of EDM for educators is a one-of-a-kind educational data repository at LearnLab [5], PSLC (Pittsburgh Science of Learning Centre), CMU (Carnegie Mellon University). They have designed some tools like:

1. DataShop
2. CTAT (Cognitive Tutor Authoring Tools)
3. TagHelper
4. ELI online data
5. TuTalk

A key goal of PSLC is to support learning scientists in providing explanations of results using, as much as possible, the same core terminology and addressing an accumulating body of precise theoretical principles of instruction. DataShop provides datasets, some of them are public and some others can be accessed on request. This is a good platform for researchers who excel in MS-Excel or have coding/programming proficiency. They have some predesigned methods like:

1. Performance profiler: a multi-purpose report, something akin to an educational researcher's Swiss Army Knife. Following measures can be viewed:
   a. Error Rate (%)
   b. Assistance Score
   c. Average Number of Incorrects
   d. Average Number of Hints
   e. Residual Error Rate Percentage (Predicted–Actual)

2. Error report: summaries of student performance by step or knowledge component, actual values students entered and the feedback they received, at-a-glance information on problem coverage – the number of students exposed to a particular step or knowledge component, the knowledge components associated with each step, and the problems that contain each knowledge component.

3. Learning Curve: visualizes changes in student performance over time.

Other than these, popular tools like WEKA, RapidMiner, SPSS, R and Matlab can also be used for the analysis. These tools have pre written code or libraries of most DM algorithms [3].

These data sets can be opened in excel and key support for attributes is included separately. These are collected with the help of Cognitive Tutor, a tool used by students in schools and colleges. These data sets need no cleaning and have no errors or empty fields. There are freely available research papers also to show what kinds of studies have already been conducted – like finding patterns and correlations. The patterns found can be used to improve classroom teaching and the tool itself. However, there are restrictions on the kind of data they can collect or questions that can be asked.

Example 1: A student has to solve an equation ax + b = c. Now there are steps in solving this equation. The tool asks the student to key-in each and every step. Student input is stored and above mentioned analysis can be performed.

Example 2: An interesting problem here could be – conduct a pre-test and a post test in mathematics class and see how students learn division and multiplication, do they realize the exceptional case of division by zero [6] and so? Do they break down the problem of multiplication in addition steps? This assesses success of students and teachers in a class.

CTAT is a tool for preparing intelligent tutors. It supports the creation of two types of tutors: example-tracing tutors, which can be created without programming but require problem-specific authoring, and cognitive tutors, which require AI programming to build a cognitive model of student problem solving but support tutoring across a range of problems.

TagHelper is a tool for applying machine learning technologies to text processing problems.

English LearnLab Data or ELI's Online Data Search System for student data collected through the various course offerings of the ELI, which is part of the University of Pittsburgh.

TuTalk is for authoring and experimenting with natural language dialogue in tutoring systems and learning research.

3. EDM FOR ADMINISTRATORS

An example of EDM for Admin is enrolment data collected by SRD (Student Registration Division) of IGNOU (Indira Gandhi...
Research Methodology for Educational Data Mining in India

National Open University. The data is stored using FoxPro and can be easily imported in MS-Excel. For this research, enrolment data of disabled student for an entire year 2009 was obtained from them in January 2010. After data cleaning some interesting patterns were obtained telling us what interested these special students. This study and its results will be discussed in sections later. Before that we will discuss about developing a research methodology for the analysis of an educational data set.

4. RESEARCH METHODOLOGY

Research methodology is a way to systematically study and solve a research problem [7]. It is necessary for the researcher to know not only the research methods/techniques which are available but also the methodology. In research methodology, the researcher selects what tools should be selected for the analysis and why taking into account all the underlying assumptions and all the criteria under consideration. This implies that the design of research methodology might differ from problem to problem. Currently, in EDM, a wide variety of DM methods are available such as prediction, clustering, relationship mining, discovery with models, and distillation of data for human judgment. An EDM researcher has to capture patterns in the data. Understanding problems of students from an educational psychology point of view is important here. An EDM research can help us in the following areas –

1. Exploration – Exploring hidden causes behind a phenomenon i.e. examining evidences. In EDM there is a lot of scope for exploring and identifying the factors affecting educators and educands. Human behavior is complex and difficult to capture. We can identify the reasons for wide disparities in performance of students in a class. How does the environment of a student affect their concentration, study methods, liking/disliking of a particular method or subject? Such answers can be partially predicted on the basis of student’s profiles. We can identify the influential factors and contributing factors from this study.

2. Description – Defining or differentiating a phenomenon from others, e.g. describing characteristics of a population or its subsets. Mining of students’ characteristics can be useful to both, administrator and teacher, in many ways. The word ‘characteristics’ here means distinguishing properties or attributes [8]. One such important attribute is ‘finance’ or ‘parental income’ which determines health, food/nutrition a family can buy and also the influence of other luxuries and exposure to technology and information. Another attribute is ‘disability’, which covers categories like physical, mental or learning disability.

3. Prediction – Identifying relationships P → Q. P is the cause of Q and Q happens after (follows) P. So P aids in Q’s prediction. For example very poor families mostly can’t afford nutritional food and necessary health care facilities. Students from such families are very likely to suffer from disability or serious ailments. This family income, history & tags/keywords selected for life style can be used to predict health conditions.

4. Explanation – Comparing two or more theories in an unbiased manner, free from authoritarianism. While dealing with enrollment data of students it can be seen that some columns contain sensitive information, like-religion, caste and creed. A researcher and the authorities under whom he or she is working should know that research being done for the betterment of mankind and not for their personal reasons or grudges. Researcher and authorities should themselves remain unbiased towards each other and the students. Similar controversial attributes are gender, marital status, and children. At some level, all these attributes have to be considered to accurately separate the most influential factors or variables. Culture and religion have a great impact on lives of students especially in countries or communities with diversity.

5. Action – Finding a solution, applying and verifying it as well. This is about usefulness of a research. An action based on the suggestive results and evaluation of its effectiveness is the best evidence to prove the vitality of such a research. In the field of education, the action based on such studies is usually lacking or delayed.

5. EVALUATION OF RESEARCH

Evaluation of research involves focusing on both sides of the coin – positive and negative, good and bad. Relevance of the research is a key decisive factor. There are following questions to be answered –

1. Who – Researchers, participants and consumers of the research. Competences and biases of the researcher play a role.

2. What – Topic and theory on which the research is based and how it is viewed by the world. EDM is a new field. There are only a handful of journals and books in this area. There are some commercial or business concerns with a focus on EDM. Most research is coming from academics. SPSS is the most popular software in this field. Books and papers with a focus on particular problems and issues are not there. EDM is considered as an important R&D area by computer science, education, social science, management and many related fields. It is also important to tie in the final objective of the research with these fields and understand the impact in each area.

3. Where – In what kind of environment research can be conducted. EDM research can be conducted in an unbiased environment. Data sets to be used here are so huge; a researcher can’t collect them individually. These databases are mostly provided by some educational body. These authorities should respond to researcher’s request if they can. Researcher should work as per rules laid down by authorities.

4. When – What is the best possible time to conduct a research. The best time is when the problem and its solution are relevant. A true research in DM is relevant at that time, even if it becomes irrelevant later.

5. Why – Other than the above, what are other motivational reasons behind doing a research. A research is important for the organizations involved it. However, we should try to fit the conclusions to the data and not the other way around even if
that does not agree with the overall picture being created by the organization.

6. How – What are the methods of conducting a research i.e. research methodology (RM). DM research can be conducted using empirical, simulated or real world data. It depends on the demand of the organization providing data and sponsorship.

6. THE SCIENTIFIC APPROACH

There are no sacred truths. A researcher must be given some freedom. There should not be too many rules as in Fuzzy Systems. But still a systematic approach has to be followed. Data is analyzed to state a conclusion, to give a theory which fits the facts (rules here) that have come out of a data. If more than one theory is possible, best fit theory is the conclusion. Tomorrow, even this final best fit theory may become invalid. Then it has to be replaced by some other theory, after a new research. Conditions, under which a research is conducted, are important. E.g. today there is stress on female education which was not there in the past centuries in India. Education budget has increased but at the same time globalization has made educational organizations look like markets. There’s lot of competition between different organization which is helping shape the current system but can also lead to negative outcomes if not managed carefully.

Other examples of this research are – when assessment data, drop out and retention is used to predict the future pathway of a student. Similarly data of leave of students, teacher and other staff can also be used for medical DM i.e. health issues of students and staff. Data of salary and perks could also be used for analysis.

Scientific research is more about observations followed by logical analysis to generalize inductively. But overgeneralization makes it nonscientific. In this research rules generated can’t be generalized to a high degree because data of only one year is being used. To achieve a high degree of generalization, data from a long duration should be used. This can also help in testing the rules generated. Then action has to be taken in the form of a pilot study. The hypothesis can then be proved that such a functional dependency of attributes is there.

6. RESULTS

In this section some of the results are presented as obtained while doing the data analysis conducted on disabled students of IGNOU who enrolled for various courses in the year 2009.

1. Most students were in their late 20s.

2. Most of them enrolled for Master of Political Sciences and paid an average fee.

3. Most students opted for English Medium.

4. Most of the students are male.

5. Most students were pass outs of past 10 years only.
Most of them are unemployed.

They are from urban areas.

7. DISCUSSION OF RESULTS

1. Result 1 is self-explanatory. Figure 3 has shape of a normal distribution as often exhibited in biological data [9]. This also resonates well with the model of our current education system where most people like to study or focus on their career value addition in their twenties or early thirties preferably.

2. Result 2 is due to the fact that these students find it easy to do humanities courses because there is no help in the form of artificial limbs & training to use them in laboratories (sciences). Travelling to various laboratories, even just to visit and have a look, is a problem as our science laboratories have no accessibility areas or equipment. It makes sense to opt for courses which require no or less resources.

3. English medium books are comparatively easily available in India. IGNOU however plans to launch courses in regional languages. More steps that can be taken are – to encourage translation of popular books/texts in all subjects to help them. And to make them accessible – brail translation, audio books (record and release), and video field tours.

4. Result 4 shows that disability is more common in males in this data sets.

5. Result 5 is self-evident. Students are in their 20s, so they have mostly passed out in recent past. Figure 6 is shape of a chi-square distribution.

6. Result 6 requires action from Governments to create accessible jobs to increase employment.

7. Result 7 indicates poor life style in urban areas. This requires looking back to our roots – Ayurveda (diet and medicine), Sports, Games, Yoga-Meditation, Arts and Music of Indian origin, simple living and high thinking created by our ancestors for us. Staying close to nature (farming, gardening) can help cleanse the body and soul of humans.

CONCLUSION

EDM research follows a very simple & straight style like any DM problem. After acquiring data, the data is cleaned. Transformations are made. After converting it in to a desirable format, a program or tool can be run on it to find patterns. Second stage is reporting the research as paper or thesis, which has to be done in a prescribed format and flow. A sample study could be finding an E-learning model based on AI techniques that fit your data.

In the mentioned two categories of EDM it was observed that there is a difference in these two researches. Since at PSLC the tools are being designed by those involved in the research, they have a very clean data. They also get to decide & choose the attributes they want. At IGNOU there may be a requirement of including a few more fields – degree of disability, monthly/annual income of the family and other personal data of family, individual/student and her assessment if possible. Assessment is not just 10th and 12th, a few degrees or a big project report. The personality and performance of every pupil have to be judged throughout the academics and the employment/career.

There are some differences in the kind of political or government hold on the data in the two universities (CMU & IGNOU) in the two democracies. But system in IGNOU is certainly an evolving one. Learning from the research is just one part. Implementation or action based on that makes the research effort worth it.

FUTURE WORK

Next step in this research is to analyze the assessment data of these students using DM techniques – like tuple analysis. Finding what determines the success of a disabled student in distance education, is the next stage of this research. This research can help us provide useful insights in the Indian education system and help us to understand the factors affecting students and improvise on them in the long run.

ACKNOWLEDGEMENTS

Authors thank the Student Registration Division at IGNOU for providing student enrollment data for this study and the team at SOCIS IGNOU for their support. The authors also acknowledges team members & mentors at PSLC Summer School 2011 (Prof. Kenneth R. Koedinger, Erik Zawadzki and Sabestian Lalle’) and Dr. Neetu Chopra (Scientist, Plextronics) for their suggestions in this work.

REFERENCES